Analytics WG - some discussion items

Dirk Duellmann
10 Sep 2014

Mandate (proposal)

- Coordinate analysis and trending of service usage data
  - typically based on days or months of collected data
    - no strict latency (<1d) or completeness requirements (<1% loss) on input data
  - with the goal of
    - getting a better understanding of a service (exploratory)
    - informing a service strategy or planning decision (hypothesis check)
    - developing & improving a predictive service model (model building)
      - using parameters extracted from real service

Mandate cont’d

- Cross group activity to
  - enable integrated studies crossing single data source / service boundaries
  - using a common base repository of prepared input data (consistency, reliable)
  - provide an exchange forum for discussion on analysis methods, tools and result validation

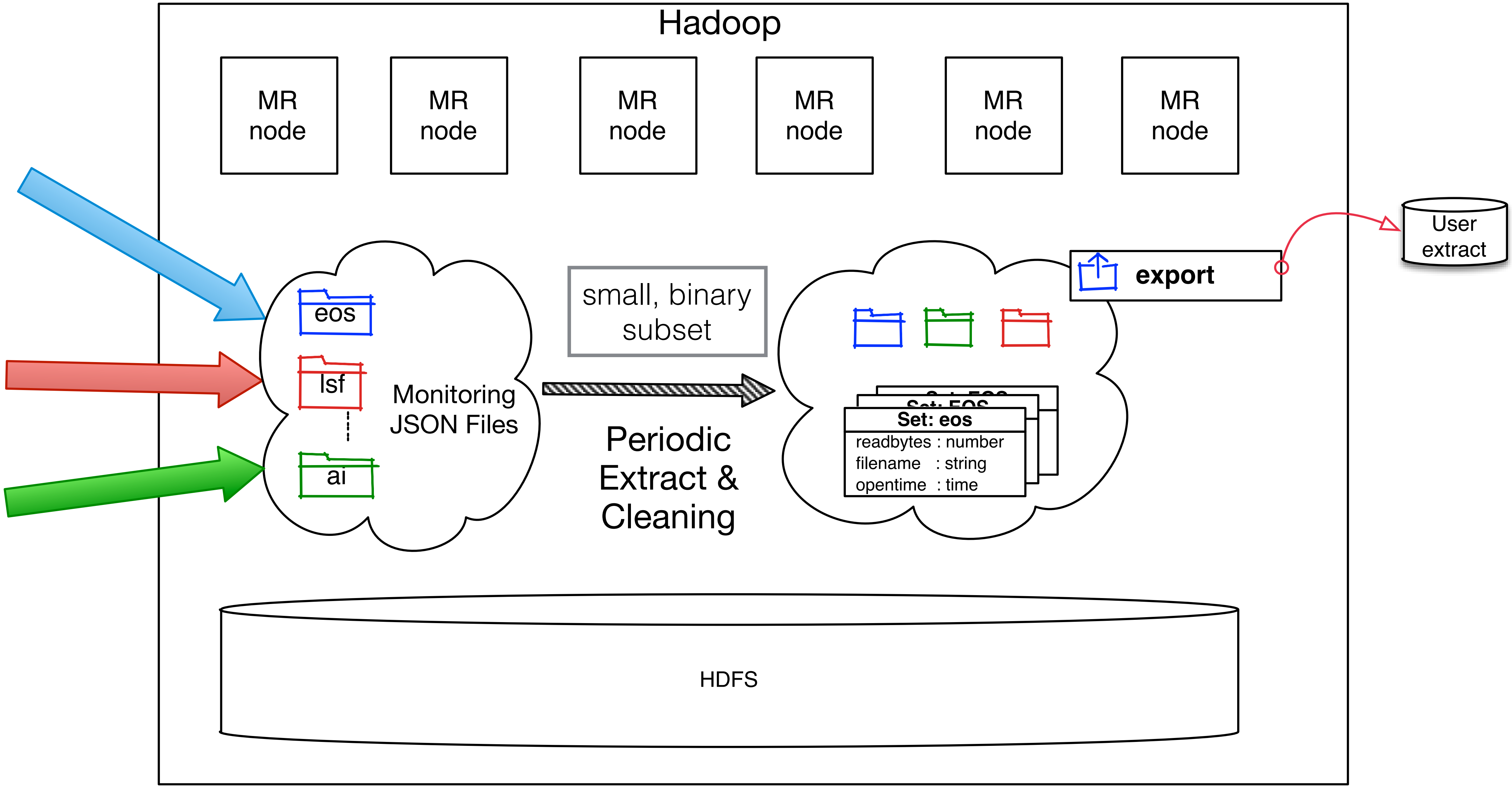
Not part of the mandate

- General purpose visualisation of direct metrics
    - done by IT monitoring / experiment dashboards
    - but analytics wg may define additional metrics / plots for inclusion in the above
- Alerting: done by IT/experiment monitoring project
        - as before
- Raw monitoring data collection
    - done by IT monitoring into hadoop for many services
    - prototype work for dashboard data in progress

Target 1: common data repository infrastructure

- Provide an repository with pre-cleaned data with export to common analysis formats and the ability to execute parallel analysis jobs close to the data
    - automated cleaning from agreed raw-data repositories (eg hadoop for IT monitoring data)
    - documents semantic, normalisation and consistency issues for known/used metrics
    - for use by developers of analytics plots, algorithms and models - not by general public

Data Collection and Analysis Repository



Hadoop

MR node

MR node

MR node

MR node

MR node

MR node

eos

lsf

ai

Monitoring JSON Files

small, binary subset

Periodic Extract & Cleaning

export

User extract

```
Set: eos
readbytes : number
filename   : string
opentime  : time
```

HDFS

Possible export scenarios

```
$ arepo -export eos,lsf –period yesterday –o eos-lsf-yd.root
# retrieved 98765 “eos” and 12456 “lsf" records from 08-09-14 (1d)
# in 5 seconds.
```

```
$ arepo -export atlas-fax:from-cern -period 01-14 -sample 1M -o cern-orig.csv
# retrieved 10**6 “atlas-fax:from-cern” records from 01-01-14 (31d)
# in 314 seconds (sample weight 0.045 of total data in period).
```

```
other output formats of interest could be:
    .sql (sql import) and .rda (R)
```

```
$ arepo -list-sets                   # show available sets and last-update time
$ arepo -list-fields eos             # show field names and short caption
$ arepo -tail lsf[1:10]              # show examples of first 10 fields
```

Target 2: linking data

- Identify (possibly missing) key-data to allow correlation (eg joining) between so-far disjoint areas
- Eg
    - CPU(box) <-> JOB(lsf) <-> storage(process)
    - IT service info <-> dashboard <-> experiment workflow
    - Hypervisor <-> VMs
- Document and improve a data model that allows to connect service areas

Target 3: Access analysis results from existing Web-portals

- Dashboard use case
  - analysis input data is existing, structured and cleaned
  - plots/algorithms are known and implemented as DB application
- Can a hadoop based repository implementation achieve similar results and user experience?
  - eg via pre-calculation of popular plots / results
  - responsive enough for integration with existing Web UI?

Some stat's topics of potential interest

- Correlation / variance analysis
  - how to quantitatively establish correlation?
- Modelling
  - parameter estimation
  - model validation
  - simulating changes
- Forecasting
  - separating seasonal effects from general trends